PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/30 — A1

(11) International Publication Number: WO 00/67160

(43) International Publication Date: 9 November 2000 (09.11.00)

(21) International Application Number: PCT/US00/12344

(22) International Filing Date: 5 May 2000 (05.05.00)

(30) Priority Data:
09/305,583 5 May 1999 (05.05.99) US

(71) Applicant *(for all designated States except US)*: EJEMONI, INC. [US/US]; 699 Mississippi Street, Suite 208, San Francisco, CA 94107 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: JEFFREY, Joel [US/US]; 606 South Washington Street, Wheaton, IL 60187 (US). BEIRNE, EoinKANGAS, Jeff

(74) Agent: DUNNING, Richard, A., Jr.; Fish & Richardson P.C., Suite 100, 2200 Sand Hill Road, Menlo Park, CA 94025 (US).

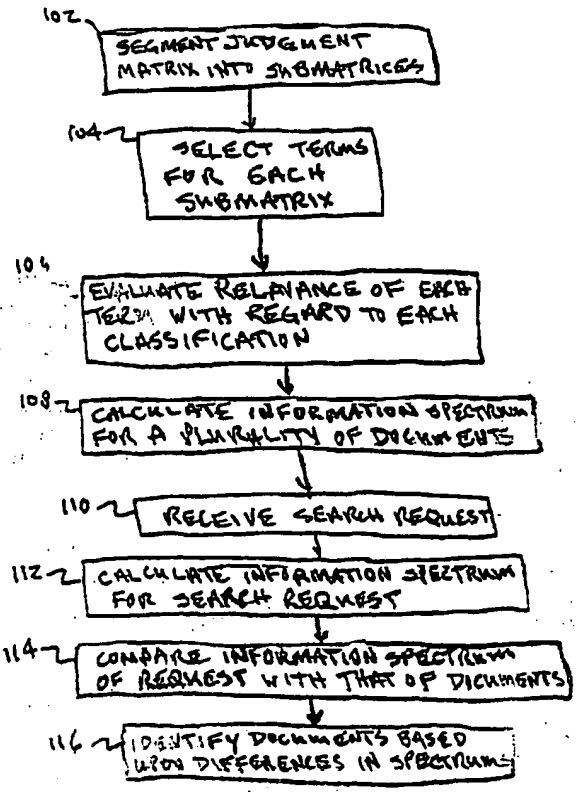
(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: WIDE-SPECTRUM INFORMATION SEARCH ENGINE

(57) Abstract

A method and computer program product for comparing documents includes segmenting a judgment matrix into a plurality of information sub-matrices where each submatrix has a plurality of classifications and a plurality of terms relevant to each classification; evaluating a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information submatrices; calculating an information spectrum for a first document based upon at least some of the plurality of terms; calculating an information spectrum for a second document based upon at least some of the plurality of terms; and identifying the second document as relevant to the first document based upon a comparison of the calculated information spectrums.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

WIDE-SPECTRUM INFORMATION SEARCH ENGINE

Field of the Invention

The field of the invention relates to document retrieval and more particularly to search engines operating within the context of a database.

Background of the Invention

Automated methods of searching databases are generally known. For example, P. G. Ossorio developed a technique for automatically measuring the subject matter relevance of documents (Ossorio, 1964, 1966, 1968, 1969). The Ossorio technique produced a quantitative measure of the relevance of the text with regard to each of a set of distinct subject matter fields. These numbers provided by the quantitative measure are the profile or information spectrum of the text. H. J. Jeffrey produced a working automatic document retrieval system using Ossorio's technique (Jeffrey, 1975, 1991). The work by Ossorio and Jeffrey showed that the technique can be used to calculate the information spectra of documents, and of requests for information, and that the spectra can be effective in retrieving documents.

However, Ossorio's technique was designed to solve a particular kind of document retrieval problem (i.e., fully automatic retrieval with complete crossindexing). As a result the technique has certain characteristics that make it unusable for information retrieval in cases in which there is a very wide range of subject matter fields, such as the Internet.

Summary

In general, in one aspect, the invention features a method and computer program product for comparing documents. It includes segmenting a judgment matrix into a plurality of information sub-matrices where each submatrix has a plurality of classifications and a plurality of terms relevant to each classification; evaluating a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information submatrices; calculating an information spectrum for a first document based upon at least some of the plurality of terms; calculating an information spectrum for a second document based upon at least some of the plurality of terms; and identifying the second document as relevant to the first document based upon a comparison of the calculated information spectrums.

Particular implementations can include one or more of the following features. Segmenting a judgment matrix can include dividing each information submatrix of the information sub-matrices into a set of columns, where each column is a classification of the information sub-matrix. It includes selecting the plurality of terms based up a relevance of each term of the plurality of terms to at least some of the classifications of the information sub-matrices. Evaluating a relevance can include assigning a numerical indicia of relevance having a range of between zero and at least two. Calculating an information spectrum for the first and second documents can include determining a log average among the:: numerical indicia of relevance of the terms of each classification. Identifying can include determining a distance between the information spectrum of the first document and the information spectrum of the second document.

It can include indicating that the first and second documents are relevant to each other. It can include using the calculated information spectrum of at least one of the first and second documents as a search request. It can include zooming in on a portion of a document information spectrum. It can include determining that the first and second documents have a wide spectra with significant content in a field F of a term. It can include measuring the first and second document using a sub-engine for field F.

In one implementation the first document can be a user interest profile describing one or more interests of a user and the second document is part of a text stream, and the implementation includes directing the first document to the user when the second document is identified as relevant to the first document.

In one implementation the first document is a profile of a user's job-related information needs, and the implementation includes denying the user access to the second document when the second document is not identified as relevant to the first document.

In one implementation the first document is a profile of objectionable content and the second document includes one or more document hosted by a web site, and the implementation includes denying access to the web site when the second document is identified as relevant to the first document.

Calculating can include calculating a depth-of-content measure for a document based on the lowest sub-matrix in which a term in the document appears; and identifying can include comparing the depth-of-content measures of the first and second documents. It can include calculating an information spectrum for a third document based upon at least some of the plurality of terms; and ranking the second and third documents in order of degree of relevance to the first document.

In one implementation the first document is a search request and the second document includes a hyperlink to a third document, and the implementation includes calculating an information spectrum for a third document when the second document is identified as relevant to the first document.

In one implementation the terms are computer instructions, the first document is a threat profile, and the second document is a set of instructions received from a source, and the implementation includes detecting a threat when the second document is identified as relevant to the first document.

In one implementation the first document is a user interest profile describing one or more interests of a user and the second document is directed to the user, and the implementation includes redirecting the second document when the second document is identified as not relevant to the first document.

In general, in one aspect, the invention features a method and computer program product for augmenting a document. It includes segmenting a judgment matrix into a plurality of information sub-matrices where each sub-matrix has a plurality of classifications and a plurality of terms relevant to each classification; evaluating a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information sub-matrices; and adding at least one of the terms to the document based on the relevance of the terms.

In general, in one aspect, the invention features a method and computer program product for evaluating a text having a plurality of terms. It includes calculating a first information spectrum for the text in a first judgment matrix having a plurality of first vector spaces arranged in a hierarchy; calculating a second information spectrum for the text in a second judgment matrix; and combining the first and second information spectra to produce an information spectrum for the text.

Particular implementations can include one or more of the following features. It can include calculating a document distance between the information spectrum for the text and a further information spectrum for a further text; and identifying the further text as relevant to the text when the document distance falls below a predefined threshold. Calculating a second information spectrum can include identifying as contributing terms any terms that appear in at least one of the first hierarchical vector spaces; identifying as associated terms any terms that are associated with one or more of the contributing terms; and calculating a linked information spectrum for the text in a second judgment matrix using each associated term. It can include calculating a document distance between the

information spectra for the text and further information spectra for a further text; and identifying the further text as relevant to the text when the document distance falls below a predefined threshold.

Brief Description of the Drawings

FIG. 1 is a block diagram of a search system in accordance with an illustrated embodiment of the invention.

FIG. 2 is a detailed block of the system of FIG. 1;

FIG. 3 is a flow chart of the system of FIG. 1;

FIG. 4 is a segmented judgment matrix used by the system of FIG. 1.

FIG. 5 depicts a primary hierarchy and a secondary hierarchy.

Detailed Description of an Illustrated Embodiment

The present invention is an information spectrum measurement engine (also referred to herein as a "wide spectrum measurement search engine") that extends prior art profiling technique to very wide ranges of subject matter, such as exhibited by the Internet, general libraries, and other broad-coverage information collections.

The primary limitation of prior art techniques is in the number of subject matter fields and the number of terms. The original techniques were based on producing a numerical matrix representing the relevance of each of a set of terms in each of a set of subject matter fields. Given S subject matter fields and T terms, each of the S x T numerical ratings must be made by a qualified practitioner in the subject matter field. The profile of a document is calculated automatically, but each item of each term profile must be produced manually; none are automatically calculated. It is an important feature of the technique that every term is numerically rated with respect to each subject matter field; each term has a manually-supplied complete information spectrum.

The current invention is distinguished from the original techniques (e.g., by Ossorio) in three ways. First, the matrix of judgment ratings of the illustrated embodiment is segmented into submatrices. Within each submatrix, a portion of the ratings are done manually, but the remainder of the matrix entries are automatically set to zero, indicating no known relevance. The information spectrum of each document is calculated from the resulting partial term spectra.

Second, the spectra of the terms may be augmented by Bayesian probabilities, which use the spectra of the documents to calculate the relevance of those terms whose ratings were

previously set to zero for each field. Known document relevance to each field is the necessary attribute for calculating Bayesian probability; the calculated document spectra provide this attribute.

Third, sub-engines may be used to "zoom in" on a subject matter, calculating the spectrum within a field. For example, a document with the terms "muon" and "Higgs boson" is measured as definitely relevant to the field of physics. A sub-engine for physics may be used to measure the information spectrum of the document relativized to physics. Sub-engines can themselves have sub-engines, providing zoom capability to as fine-grained a level as there are recognized subject matter fields.

FIG. 1 is a block diagram of a searching system 10, generally in accordance with an illustrated embodiment of the invention. As may be seen from FIG. 1, a central processing unit (CPU)16 (and included search engine 28) may receive information and commands from any number of sources.

FIG. 2 is a block diagram, which shows processing blocks that may operate from within the search engine 28. FIG. 3 depicts processing steps used by the engine 28 of FIG. 2. Reference shall be made to FIGS. 2 and 3 as appropriate to an understanding of the invention.

Documents may be received from a document source 22 and processed directly, or stored in a database 18. Alternatively, the CPU 16 may recover documents through the Internet from other locations (e.g., other databases) and process the documents directly or, again, store them in the database 18.

It should be noted at this point that only an information spectrum of a document need be stored in the database 18 for searching purposes. The requirement that the database only store an information spectrum of a document significantly reduces the storage requirements of the database 18. For recovery purposes, a hyperlink may be stored in the database 18 along with the information spectrum which, in turn, may lead a user to a database containing the original document.

A system administrator 24 may enter information classifications or terms relevant to classifications. An expert in one or more classifications may evaluate terms relative to any classifications entered by the system administrator 24.

Alternatively, the CPU 16 may receive search requests from a user operating through a local or remotely located terminal (12)(the user and terminal will hereinafter together be referred to as "user 12"). The user 12 may access the CPU 12 from a remote location through the Internet 14 or locally through a direct connection 30.

Turning now to operation of the CPU 16, an explanation will be provided of the steps

used by the search engine 28 in accomplishing the improved search method. While the steps used will be described with some generality, it should be understood that the steps described are embodied in the programming steps practiced by the CPU 16.

As, a first step, a method of constructing a specific type of judgment matrix will be discussed. Following the discussion of the construction of the judgment matrix, is a discussion of how the judgment matrix is used.

The wide spectrum information measurement search engine is an advance of-prior techniques in two aspects. First, the necessity of manually producing an entire judgment rating matrix is eliminated by segmenting the judgment matrix. Second, sorting of the results is eliminated. Each of these aspects will be discussed in more detail below.

In general, a judgment matrix (FIG. 4)is made up of a number of rows (with a term t_a associated with each row)and a number of columns (with a classification F_b associated with each column). The classifications refer to subject matter classifications. The terms are words that may be used to describe various aspects of each classification.

Under previously used methods, a set of subject matter fields were selected for creation of the judgment matrix. Any set of fields was permissible, so long as the set was inclusive of the entire information spectrum. No provision was made for overlap of fields or for the effects of any possible relationships between the fields.

For each field of the prior method, a set of documents were selected. The documents are selected by a competent person as being clearly within that field. A set of terms were selected for each of the prior method. The selected terms are words and phrases, taken from the documents for that field that are recognizable to persons competent to make the judgment as being at least tangentially or vaguely related to that field.

Putting the subject matter fields as column headings and the terms as row labels, one has an empty table. From a set of competent human judges, ratings are collected of the degree to which each term is relevant to each field. These ratings of the prior method differ from the use of more customary subject field codes or topic tags in two ways. First, they are not simply a "checking off" that a term is part of a field. The degree of relevance, or importance, of the term is part of the rating.

Second, the rating is numerical. The judges use the following scale in making ratings. If the term is irrelevant, the rating is zero. If the term is tangentially or vaguely related, the rating may be one or two. If the term is peripherally relevant, the rating may be three or four. If the term is definitely relevant and something clearly within the field, the rating may be five or six. Finally, if the term is a highly significant concept within the

field, the rating may be seven or eight. In each category, the higher number is used to indicate greater relevance or significance.

Relevance is quantified in a pragmatically useful way; the non-binary nature of relevance is represented and used; and more importantly the ratings are not statistical in nature. There is no relationship, in represented by a term and the statistics of its occurrence in a corpus of text. Schroedinger's equation, for example, is a central and crucial concept in quantum mechanics, but a text or article on quantum mechanics may have few or no actual instances of the term "Schroedinger's equation." Just as, in information theory, the information value of a signal cannot be determined from the characteristics of the signal, but can only be found from the context (specifically, the possible values of the signal), relevance of a term to a field refers to the place the concept has in the practices that comprise that subject matter field, that is, to how the concept is. used in the field. The fundamental advance of prior methods was to devise a technique for representing; in computer-processable form, information about terms that is not derivable by any statistical, mathematical, or algorithmic process. As we shall see, statistical and other formal methods may be used to augment a set of term relevance ratings to provide an initial set, but that initial set is not statistical. This has significant implication for the novelty of the new technique of embodiments illustrated below.

The matrix of judgment values may in general have considerable overlap and redundant information, because the fields themselves were originally selected with no thought to their relationships. If we were to view the fields, the columns of the matrix, as a mathematical basis for each of the terms' vectors, the mathematical statement of this situation is that the basis is not minimal and is not orthogonal. Overlapping and redundant fields seriously harm the use of the vectors, as will become apparent shortly. To use the vectors, an orthogonal basis for the set of vectors is highly desirable. This basis is found by factor-analyzing the judgment data matrix. If there is prior reason to know that the original dimensions are conceptually independent, this step may be skipped.

The measurable common factors, together with the unique factors, resulting from the factor analysis provide the basis of the vector space. Each common factor is divided into two factors: those with significant loadings (over 0.7), and with significant negative loading (less than -0.7). If the rating step described above is skipped, this is equivalent to counting each original field as a unique factor, with loading 1.0.

The information profile, or spectrum, of each term may be calculated by averaging the ratings of the term on the fields that make up each basis vector, weighting the average by the cube of the loading of the field on the vector. For example, if basis vector 1 is

comprised of fields 1, 3, and 17, with loadings of 0.8, 0.9 and 0.85, respectively, and term is rated 4 in field 1, 6 in field 3 and 8 in field 17, then component 1 of term t's profile is given by: (0.83*4+0.93*6+0.853*8)/(0.83+0.93+0.853).

The information spectrum of each document is now calculated by combining the term profiles of all terms found in the document. In calculating the document spectrum, due to the orthogonality of the basis vectors, only the 1st component of the term spectra contribute to the 1st component of the document spectrum, only the 2nd components contribute to the 2nd component, and so forth. The most recent work (Jeffrey, 1991)used a log-average of the component values.

Retrieval is accomplished by scanning a user request for known terms, calculating the spectrum of the profile of the request (e.g., as for a document), and calculating distance from the request spectrum to each document spectrum. Any distance measure may be used; Ossorio and Jeffrey used Euclidean distance; Jeffrey also used the Manhatten distance. Experiments by Jeffrey with other distance measuring techniques demonstrate that change of distance measurement techniques does not result in a significant change in the procedure. Since the spectra represent subject matter content, spectra that are similar numerically have similar subject matter content. Retrieval of documents in the past has proceeded by sorting all documents in order of closeness of spectra to the request spectra, and returning documents to the user in order of closest first.

The technique is described herein in terms of a search for documents of a predetermined subject matter. However, Ossorio showed that the overall technique can be used to measure the spectrum of several kinds of information, such as attributes, categories, significant dimension of variation and means-end (Ossorio, 1966, 1969). In a medical context, J. D. Johannes showed that the spectrum can be the diagnostic indications of a set of patient signs and symptoms (Johannes, 1974).

Difficulties arise in attempting to extend the basic techniques of the prior art to situations in which there is a large number of subject matter fields (or categories, types of content, etc.). The most serious of these is the amount of time necessary to complete a rating matrix. Ossorio constructed measurement systems from 60 subject matter fields and 1548 terms, a total of approximately 93,000 individual ratings. In Jeffrey's document retrieval system 62 subject matter fields and subfields were selected from the area of Computer Science. To describe the subject matter, 800 terms were used which required approximately 48,000 ratings. In that work, it was found that approximately 1,000 ratings can be done per hour. A 20-field, lO,OOO-term matrix could thus be constructed in about 200 hours.

However, in the case of the Internet, for example, a very conservative estimate of the number of subject matter fields necessary to cover all subject matter on the Internet would be at least 1,000, and would require at least 100,000 terms. This rating matrix would take 100,000 hours, or 50 person-years, to construct.

The second difficulty or prior methods arise in retrieving individual documents from a very large collection of documents. The distance from each document to the request must be calculated, either by calculating each document distance individually or by applying an automatic clustering technique to the text of the documents or to their spectra. Calculating this distance, for SO-component spectra, takes approximately 130 seconds for l, OOO, OOO documents on a PC with a 300 MHz processor. However, the basic retrieval technique requires sorting these distances, to retrieve the most similar document first. The fastest possible sorting algorithm requires time proportional to N*log(N), where N is the number of items to be sorted. As a result, sorting l,OOO,OOO documents requires 3,000 times as long as sorting 1,000 documents (not 1,000 times). (By way of comparison, the Windows DOS sort command requires 1 min 15 seconds to sort 900,000 numbers, on a 300 MHz PC with 64 megabytes of RAM. Sorting 10,000,000 would therefore take 10*4=40 times as long.) These two difficulties make document retrieval by information spectrum impractical for very large databases, such as the Internet.

The wide-spectrum information measurement engine 10 of FIG. 1 is a significant advance over prior systems in two respects. First, the necessity of manually producing an entire judgment rating matrix is eliminated, by segmenting the judgments. Second, the need for sorting is eliminated.

The judgment matrix of the illustrated embodiment of FIG. 4 is developed as follows. First, the columns of the matrix (e.g., the subject matter fields, when the matrix represents subject matter relevance) may be segmented 102 into groups G_1, G_2 . . . G_n. Each group may be divided to include a number of classifications F_1 to F_a. For each group G_i, a set of terms t_1 to t_b are selected 104 for each of the fields. For each field, a set of documents clearly within that field may be selected by a competent person in that field. A set of terms are selected for each field. These terms are words and phrases, taken from the documents for that field, which are recognizable by persons competent to related to the field.

For each group G_i, and the terms for that group, ratings are obtained for each of the terms with respect to each of the fields. The subject matter fields are placed into the segmented judgment matrix of FIG. 4 as column headings and the terms as row labels.

From a set of competent human judges, ratings are gathered and evaluated 106 of the degree to which each term is relevant to each field. The judges may use the following scale in making ratings: if the term is irrelevant, the rating is zero; if the term is tangentially or vaguely related, the rating may be one or two; if the term is peripherally relevant, the rating may be three or four; if the term is definitely relevant and something clearly within the field, the rating may be five or six; and if the term is a highly significant concept within the field, the rating may be seven or eight. In each category, the higher number is used to indicate greater relevance or significance. However, each term is rated only with respect to each of the fields which make up the group. All other matrix entries are set to zero. The result of this procedure is a matrix of entries as illustrated by FIG. 4.

Under the illustrated embodiment, the procedure segments 102 the overall ratings matrix into a disjoint set of smaller submatrices such that every term is rated with respect to the fields of one submatrix. Further, each submatrix has a set of terms which represents its content.

As with prior methods, the matrix of judgment values selected may in general have considerable overlap and redundant information, because the fields themselves were originally selected with no thought to their relationships. If the columns of the matrix are to provide a mathematical basis for each of the terms' vectors, the mathematical statement of this situation should be minimal and orthogonal. To use the vectors, an orthogonal basis for the set of vectors is desirable. This basis is found by factor-analyzing the judgment data matrix.

Factor analysis re-distributes the original evaluation data provided by the judges. The original columns (i.e., classifications)(now called subject matter fields)are grouped together into common factors. The number which relates the original fields to the groups is called the factor loading. The output of the factor analysis is a set of factor loadings. The set of factor loadings represent the angle between the original evaluation data and the factor analyzed evaluation data. Factor analysis may be accomplished using any of a number of commonly available software packages provided for such purpose.

The information spectrum for each term of the entire matrix may now evaluated 106. The information profile, or spectrum, of each term is calculated, as above, by averaging the ratings of the term on the fields that make up each basis vector, weighting the average by the cube of the loading of the field on the vector. However, the spectrum components for a term not rated with respect to some field F_k is automatically zero, since that term's rating on the field was automatically set to zero.

The information of each document in the collection of documents is now calculated 108 in a first information spectrum calculator 66, using the term spectrum discussed above. The collection is now ready for use, in any application in which it is useful to have the information spectrum, such as for retrieval in response to a-user request.

Retrieval is accomplished as follows. A received request 110 is scanned for known terms and its information spectrum calculated 112 in a second information spectrum calculator 68. An information spectrum of the request and documents may then be compared 114 in a comparator 62. Documents may be identified 116 and retrieved based upon Euclidean distance of the document spectrum from the requested spectrum.

Each segment G-G_n is, in effect, a basic information spectrum measurement engine. Ossorio's results, confirmed by Jeffrey, showed that the structure of the basic measurement engine is stable if there are a minimum of approximately 6 term per field. This allows the user to calculate the effectiveness of the segmentation process. If the user is constructing a wide spectrum information measurement engine on, for example, 1080 fields, using ratio of 6 terms per field, the user has an overall matrix of 1,000 field by 6,000 terms or 6,000, OOO entries. At 1,000 ratings per hour, this matrix would require 6,000 hours (three person years of manual effort). However, by segmenting the matrix into groups of 50 fields, each segment would require 50 fields x 300 terms, or 1.500 ratings, which requires 1.5 hours of effort. To cover the 1,000 fields requires 20 such segments, resulting in a total effort of 30 person-hours, or 5% of the effort to manually fill out the entire matrix.

This reduction in effort is not without cost. It was noted in the discussion of the basic technique that the basic technique was devised in order to produce fully automatic and completely cross-indexed document retrieval. By setting large portions of the ratings matrix to zero, some cross-indexing information is lost. The value of this cross-indexing in retrieving all relevant documents is restored by the person doing the search. The searcher may need to use the retrieved documents to continue the search, by having the measurement engine based retrieval system search a second time for documents similar in spectrum to one or more documents already retrieved. The user may receive documents on his terminal 12 downloaded from a document reviewer 64 and select a document. The document selected is treated like a request (i.e., terms are identified in a term extractor 60), and other document with similar spectra are retrieved. By "pasting" portion of several documents into a request (using the WORD. facilities of the terminal 12), requests of arbitrary size and scope can be composed.

To illustrate this point, consider a spectrum measurement engine covering fields of history and of medicine. A figure of great importance in English history is King Henry VIII. Henry VIII is known to have had syphilis. A searcher wants to find documents that discuss Henry's medical condition. She requests documents on King Henry VIII. This name is rated as highly relevant (7 or 8)to English history, relevant (5 or 6)to the field of history in general, and zero (by default)with respect to the field of medicine in general and the subfield of sexually transmitted disease (STD). Therefore the searcher's request, containing only terms relevant to English history, will have a measured spectrum high on English history and very low or zero on medicine. Documents with similar spectra will be returned. Since "Henry VIII" was not rated with respect to medical fields, documents with high medical and STD content will not be returned. However, since documents on Henry VIII will have similar spectra, they will be returned to the searcher, and some of these will mention syphilis. The searcher selects one of these documents, or a portion of it (e.g., one with a greater medical content) and requests documents similar to that one. This second retrieval produces documents with much higher medical and STD content and much less history content.

While six terms per field may suffice to construct a. basic spectrum measurement engine, or a segment of one, calculating the information spectra of a large collection of documents in a field requires a much larger vocabulary, for document spectra are calculated solely on the basis of the spectra of terms found in documents. While 300 terms will suffice to construct a basic measurement engine for 50 fields, several thousand terms may be needed to cover the usage in documents. In Ossorio's original work, he found over 1,500 terms in a corpus of 36 documents. For 50 fields, 1,000 terms can be manually rated, as was done in Jeffrey's work, but 10,000 terms would require 500 hours.

Once a basic information spectrum measurement engine is complete, the following procedure can be used to calculate the term spectra for additional terms. First, the document profiles of the document collection are calculated using the initial segmented wide-spectrum information measurement engine described above.

For each new term, the following steps may be followed. For each orthogonal dimension, d, count the number of occurrences of the term in all documents in the collection. Count the number of occurrences of the term in all documents considered definitely relevant to dimension d. (Typically this will be defined by having a rating of greater than or equal to 5.0 on dimension d.) However, a lower threshold can be used if the engine designer desires to make the engine more likely to rate a document relevant to dimension d on the basis of the occurrence of terms less connected to dimension d.

The probability that a document with this term had dimension d content is given by the Bayesian probability formula as follows:

P(d(term t)=p(d &term t)/p(term t)

The probability p(d &term t)and p(term t)are given as follows. First p(d &term t)are made equal to the number of documents with term t that are relevant to d divided by the number of documents in the collection. Second, p(term t)is made equal to the number of documents in the collection.

For example, given a collection of 10,000 documents, with 1,000 relevant to physics (i.e., rated 25.0 on the dimension of physics), the term "muon" is found in 100 physics documents and 20 non-physics documents.

P(physics &muon)=100/10,000=0.01

P(mu0n)=(100+20)/10,000=0.012.

Therefore,

P(muon|physics)=0.01/0.012=0.831.

Since the spectra are normalized at 8.0, (8,0=highest degree of relevance), this probability is multiplied by 8, to yield a calculated (not human supplied) relevance of 6.6.

When a term is found in a small number of documents, this procedure is not reliable, due to small sample size. Variants of the process are: 1) do not calculate a relevance value for a term appearing in less than a minimal number of documents relevant to dimension d, or (2) set a heuristically determined value of 1.0 to 3.0 for a calculated relevance in this case.

In Ossorio's original work (1964), he noted the need for a way to "zoom in" on a portion of a document's information spectrum. The system 10 provides that capability.

Suppose that field F (or, in the case of an engine measuring another type of content such as an attribute-measurement engine, component F of the measured content spectrum) is represented in a wide-spectrum engine, and that no subfield of F are represented in the wide-spectrum engine. Further suppose that a secondary engine, covering the subfields of field F has been constructed, either by the basic spectrum measurement technique or the wide-spectrum technique described above. For example, F could be Computer Science and the sub-engine the Computer Science subfield engine devised by Jeffry. A document and request are both determined to have wide spectra with significant content of type F (i.e., over a threshold value, typically 5.0). If the difference in values on component F is "w", w is the contribution of the difference in content F to the distance between the

document and request wide spectra. If both document and request had no other nonzero components in their spectra. The Euclidean distance between them would be w. However, both document and request are now measured by the sub-engine for field F, and it is found that he Euclidean distance of their spectra within field F is f. If the engine for field F has N orthogonal components, and each spectrum has a value of at least 5 on at least 1 component of F, the maximum distance D_{max} between the 2 spectra is determined as follows:

$$D_{max},=\text{sqrt}((N-1)*8^2+(8-5)^2)=\text{sqrt}((N-1)*64+9).$$

The difference w between the document and request spectra on field F is replaced with a value equal to $F*w)/D_{max}$.

Thus, if the document and request are as far apart as possible within F, the difference in their wide spectra used to compute their retrieval distance remains almost the same. However, if they are quite close within F, their wide spectra retrieval distance is correspondingly reduced. If F is the only significant content for the document and request, the document is thus retrieved much earlier in the sequence of retrieved documents as is appropriate.

In general, the zoom-in procedure is recursive, as subfields of F (or sub-types of content type F) may themselves have sub-subfields and these may be represented by sub-subfield spectrum measurement engines. The recursion is limited only by the particular sub-engines implemented.

The subject of elimination of sorting will be discussed next. Retrieval with the basic spectrum measurement engine depends on sorting the documents by their distance from the request. Suppose three documents are found whose spectra are at distances 3.0, 5.0 and 3.5 from the request spectrum. If retrieval is to .be in order of most-relevant first, the documents must be sorted in order of distance that their spectra are from the request spectrum. However, the same is not true if the distances are 3.0, 3.02 and 3.01. The fundamental concept of relevance is that of practical use by a person. The concept is a pragmatic (not numerical) one. Accordingly, although these distances are numerically out of order, the differences in distance are not significant. This is due to the fact that the original ratings, upon which all numerical calculations are based, are integers from 0 to 8. Using standard scientific rules of precision, differences of 0.1 are significant, but differences of less than 0.1 are not. If effect, the space of all information spectra derived from the basis rating procedure is quantized. Therefore the current invention alters the basic measurement engine retrieval procedure as follows.

First, all documents at distance d are placed in a "bucket" whose number is the integer part of d/0.1. Thus, documents at distance 0.0 to 0.1 are placed in bucket 1, those from 0.1 to 0.2 in bucket 2, etc. The maximum number of buckets is given by the maximum distance two spectra can have, if they have N components: 8*sqrt(N).

Second, all documents in bucket 1 are retrieved for the user, then all in bucket 2, etc. While documents in the same bucket may have numerically different distances, the distances are not meaningful, and therefore retrieval in order of relevance is not violated.

The effect of this procedure is to eliminate the sorting step from retrieval. As noted above, this is a very-significant savings in time for retrieval, of particular importance for searching large document collections, such as the Internet. The novelty of this advance is the recognition that the information spectrum space is quantized. This is what allows the bucket technique without degradation in retrieval performance.

The information-spectrum measurement engine 10 differs from prior techniques in that it measures the subject matter relevance (or other type of content) of text, quantitatively. The spectrum is a normalized numerical measure of the amount of each type of content the text contains. This distinguishes it from all methods, processes, and systems that perform calculations to associate a set of subject matter fields by name, a set of words, or a network of words linked by named relationships, such as is done with a semantic network and from systems that produce a vector of words, attribute labels, a subject matter field labels or decodes, or other names, labels or representative tokens. Further, it does not attempt to "understand" the language of the text in any way other than to measure its information spectrum.

As described above, sub-engines may be used to "zoom in" on a subject matter, calculating the spectrum within a field. Sub-engines can themselves have sub-engines, providing zoom capability, thereby establishing a hierarchy of vector spaces. Each sub-engine corresponds to a vector space in the hierarchy. Thus, content is represented in a set of hierarchically organized vector spaces, referred to herein simply as a "hierarchy."

Now several implementations are described in detail.

1. IS-based text stream filter.

One implementation provides an information-spectrum-based text stream filter.

"Filtering" is the name of the process of directing items from an incoming text stream to a user based on a defined interest profile. In conventional filtering systems, this

user interest profile is a set of terms (words and phrases). The profile can be produced automatically or manually. The terms can have associated numerical weights, based on various statistical measures of term frequency. An incoming item is directed to a specific user if a large enough proportion of the terms in the item match the terms in the user's interest profile. This process is recognized as difficult to do well and error prone by all practitioners in the field.

The IS-based text stream filter uses an information spectrum (IS) to describe the conceptual content of interest to a user. As described above, an IS-based system includes a set of vector spaces, hierarchically organized. The top level space represents the range of content, at the broadest level. The spaces at the next level down in the hierarchy represent the content, at the next level of detail, of each of the fields from the top level, and so on for all levels of the hierarchy. An IS for a text is a set of vectors, one from each of a set of spaces in an IS hierarchy. This set represents the aggregate conceptual content of the text, at every level of detail. The IS of a user's interest profile represents the aggregate conceptual content of the user's information interest, at all levels of detail

A user's IS profile is calculated from at least one of the following: a user's statement of interests, the information spectra of the user's requests for information, and the information spectra of documents the user has indicated an interest in, by reading them, viewing them on their computer screen, saving them, or otherwise indicating interest. These spectra are combined into a single information spectrum via a mathematical process that combines a set of spectra into one, as described above or by other formulae. The user need not be a person; a set of documents accepted by any agent, whether human or automated device, provide a profile of that agent's interest in the fashion described. A document is directed to a user when the distance between the IS of the document and the IS of the user profile is less that a predetermined threshold value. This threshold value may be selected and adjusted by the user or an administrator.

This distance is calculated as described above. Terms and term similarity are not used in any way in deciding whether a document is to be directed to a user, and therefore the device is entirely different in operating principle from all existing devices for this purpose.

Since distance between the document and the user profile represents numerically the similarity of the content of the document and the content of the user interest profile, the method solves the problems associated with previous filtering systems. No term statistical information is used. Similarity of terms in a user profile and a document

correlates poorly with content similarity, and therefore term similarity filters produce results that correlate poorly with the user information need.

2. IS-based security monitor.

One implementation provides an IS-based security monitor.

It is common in high-security organizations that deal with information, such as various government agencies, financial institutions, and others, to arrange the work of the organization in a "compartmentalized" fashion, so that a person whose work involves Subject A knows as little as possible about Subject B. The person working in Subject A may issue requests for information and in return receive documents about Subject A. If the Subject A person is known to be making requests for information about Subject B, or to be receiving documents about Subject B, it is cause for inquiry by the members of the organization responsible for security in the organization.

It is not practical to have a person monitor the requests and documents from a multiplicity of workers, for two reasons. First, the volume of requests and documents is great. Second, recognizing when a request is about Subject B requires that the human monitor know about Subject B and that Subject B exists. Allowing a monitor to know about various subjects, and that they exist, is a serious violation of security.

A computer with appropriate software can perform this monitoring function because it can process the volume of requests and documents. In addition, having a record in the computer that Subject B exists and what it is about is not in itself a violation of security. The IS-based security monitor performs this function by determining the spectra of each worker's information request and received documents and measuring the distance between the person's known job-related information needs and the requests and documents the worker is receiving. The monitor can then (1) Refuse to deliver any document too distant from the worker's "need to know", (2) Alert an internal security officer, or (3) Take other action the organization deems appropriate. These actions may be triggered automatically, with or without the worker's knowledge.

The monitor can measure similarity of non-text inputs as well, such as the degree to which a given set of computer commands is associated with a task that is part of a worker's area of responsibility. This implementation ensures that a worker's commands to a computer or other automated system are consistent with his/her job responsibilities.

This functionality cannot be accomplished by any conventional technique based on occurrence terms, because term occurrence similarity cannot be used to measure

aggregate content similarity, the central feature this implementation, by any known means.

3. IS-based contract tracker.

One implementation provides an IS-based contract tracker.

It is common in large organizations to have need to find a contract, or portion of a contract, for goods or services that are related by having similar content to another contract or portion of one. This task is quite similar to the task finding a document in response to a request. This device functions by measuring the IS of a cited contract or portion of one and finding all contracts or portions with similar spectra. The spectra measurement and distance calculation are implemented as described above.

4. IS-based objectionable content filter.

One implementation provides an IS-based objectionable content filter.

A number of web sites supply content that is found objectionable by a significant number of people. Such content can include material that is sexually explicit, that is degrading or abusive to a particular social group, and that advocates particular political views, such as Nazism, and the like.

There is currently no automated way to recognize such content and prevent minors from receiving it, either at their request or accidentally. A number of such devices attempt to do so, but all are based on scanning documents such as Web pages for particular words or phrases, and refusing to load such a documents. All such devices either block a significant number of documents that are not objectionable or fail to block a significant number that are objectionable. These problems are not solvable by any known technique, for two reasons. First, while some terms are sufficiently objectionable that their simple presence merits blocking the document, many others are not. For example, a filter that blocks pages with the word "ass" will block certain entirely acceptable biblical passages. Second, certain pages, including many sexually explicit ones, contain little or no text, and thus cannot be filtered using simple term matching.

The IS-based content filter works on an entirely different principle. The filter calculates the IS of an entire web site, using one or more documents on that site that can be profiled, thereby producing an aggregate content IS for the site. A document is then filtered out or allowed to load on the basis of the IS of its origin site, whether or not it contains any particular words or phrases (or none at all). A user, such as a parent or

administrator, defines a profile of objectionable content, either by description or by specifying examples of pages or sites. Pages with content, or from sites, with an IS too close to the IS defining objectionable content provoke a user-defined level of response, such as refusal of the page, alerting an administrator or parent, and the like.

The filter can also be used by those desiring to locate web sites providing a particular type of content, such as law enforcement agencies or groups monitoring ethnically degrading or abusive content.

5. Multiple-aspect IS-based search engine.

One implementation provides a multiple-aspect IS-based search engine. This implementation uses two or more different types of content, each represented by a hierarchy. For example, the system may have a subject matter field hierarchy and a geographic information hierarchy. One of the hierarchies is designated the primary hierarchy. For example, the subject matter hierarchy is designated as the primary hierarchy.

The hierarchies are used in parallel. A document has an IS in each hierarchy of vector spaces. When a request is received, its IS is calculated in each hierarchy. The distance from the request spectrum to each document spectrum, referred to herein as the "document-request distance," is then calculated in each hierarchy, as described above. The document-request distances in the hierarchies are then combined into a single document-request distance. For example, the distance D can be found by:

$$D = \sqrt{D_0^2 + c_1 D_1^2 + c_2 D_2^2 + \ldots + c_n D_n^2}$$

where D_0 is the distance in the primary hierarchy, the D_i, $1<=i<=n$, are the distances in the additional hierarchies and each c_i is a numerical factor giving the relative importance of that type of content relative to the primary content.

6. Linked-aspect IS-based search engine.

One implementation provides a linked-aspect IS-based search engine.

In the multiple-aspect IS-based search engine described above, the IS of an item of text is calculated in each hierarchy with no consideration given to the relationships between the terms that are in each of the respective hierarchies. The linked-aspect IS-

based search engine uses certain of the additional hierarchies in an additional way. The content of the primary and additional hierarchies is linked as follows:

a) The IS of the text in the content hierarchy considered primary is calculated, in the manner described above.

b) For each vector space in the primary hierarchy, all terms with content in that space are identified. These are the "contributing terms" in this vector space.

c) For each vector space in the primary hierarchy, terms from a second hierarchy that are associated with one or more contributing terms are identified.

d) These associated terms are used to calculate the IS of the text in the secondary hierarchy.

The secondary hierarchy may be linked to the entire primary hierarchy or to any portion of it. The association between the contributing terms and the secondary terms may be any relationship that can be automatically recognized: proximity, syntactic (such as subject-verb, noun-adjective, etc.), or any relationship determinable by use of IS technology for recognizing relationships not otherwise recognizable algorithmically. For example, a separate IS space can recognize emotional content and connotation of words from the primary space. The linked secondary terms are then those secondary space terms that indicate these emotions. The distance between any two texts is then the distance including the primary and linked IS, in each space of the hierarchy. A numerical multiplier provides relative weight of the linked secondary hierarchy content.

A more detailed description, and an example, is now described with reference to FIG. 5. In general, the system includes a primary hierarchy and one or more secondary hierarchies. In this example, a primary hierarchy describes subject matter content, and a secondary hierarchy describes geographical information. The primary hierarchy includes single field 502 of Finance, with a sub-field 504 of Macroeconomics and a sub-field 506 of Investing. The secondary hierarchy includes the field 508 of Europe, with a sub-field 510 of France and a sub-field 512 of Germany. Each field of each hierarchy contains one or more terms, as shown in FIG. 5.

One implementation processes an item of text as follows.

a) The IS of the text is calculated in the top space of the primary hierarchy, using all the recognized terms in the text. For example, if the text is "I am interested the influence of German inflation on French bonds", the terms "inflation" and "bonds" are recognized.

b) Find all terms in the text that are terms in the top space of the primary hierarchy. These are the contributing terms in the top primary space.

c) Find all terms in the text that are terms from the secondary hierarchy that are associated with any of the contributing terms in the top primary space. In this example, the associated terms are "France" and "Germany".

d) Calculate the IS of the text in the secondary hierarchy, using all the associated terms found in step d), in the manner described above. This is the linked secondary IS.

e) Descend the primary hierarchy, in the manner described above. At each space of the primary hierarchy:

i) Calculate the IS of the text using the terms in the text that are in the space. In this example, in the Macroeconomics space, there is only one term, "inflation", and in the Investing space there is only one term, "bonds".

ii) Find the terms of the text that contribute to the IS in this space. In this example, in Macroeconomics the sole term is "inflation", and in Investing the sole term is "bonds"

iii) Find all terms of the secondary hierarchy associated with terms of the primary hierarchy in this space. In this example, there is only one term in the Macroenomics space, "inflation", and there is only one term of the secondary space associated with "inflation": "Germany". In the Investing space, the only term of the text is "bonds", and the only associated secondary term is "France".

iv) Calculate the IS of the text in the secondary hierarchy, using only the associated secondary terms. In this example, for the Macroenomics space this calculation produces an IS in the secondary hierarchy based only on the term "Germany", and for the Investing space the calculation produces an IS in the secondary hierarchy based only on the term "France".

f) In calculating the similarity between two texts, for example between a request and a document, use the linked ISs of the request and document, weighting the distance between the linked IS of the request and document with a numerical factor representing the relative importance of the type of content represented by the linked secondary hierarchy.

In this example, suppose that the distance between the primary IS between the document and request in the top space is 2.1, and the distance between the request and document IS in the Europe space is 1.2. Then the combined distance between document and request, or more generally between any two items that have linked ISs, incorporating

both the primary IS and linked secondary IS, may be calculated by a formula, which may include a relative weighting factor for one or more of the secondary ISs. For example:

$$D = \sqrt{2.1^2 + c1.2^2}$$

where c is the weighting factor, specified by the system designer.

Example of the use of this engine: The following request is received: "Information about rising costs and falling profits in the automotive industry". This request has two types of content: subject matter (business topics including profitability, operating costs, etc.), and activity content: Something is rising and something is falling. The linked-multiple-aspect engine retrieves articles about rising costs and falling profits, but not those about rising profits and falling costs, because the distance from the latter documents to the request is greater than the similarity threshold for retrieval.

7. **IS-based user profiler.**

One implementation provides a IS-based personal profiler.

It is often desirable to have a profile of a person's interests. Conventional methods create a personal profile by gathering personal and demographic data about the person, and by compiling a list of terms from queries the person has submitted, and documents the person has viewed.

The IS-based personal profiler calculates an aggregate content profile from either or both of (1) a person's information requests and (2) the documents they view. The requests and documents yield a set of spectra. These spectra are combined mathematically, in the manner described above, into an aggregate content spectrum. This spectrum profiles the user in all facets of content implemented on the particular system.

This IS profile can be automatically updated according to new requests and documents. The various inputs of new information can be weighted so as to place greater emphasis on some information. For example, recent requests and documents can be weighted more heavily than older ones. In addition, a person can proactively alter his profile to add and remove interests.

IS-based interest profiles can be used for a variety of purposes, including: (1) to automatically supply a user with information and offerings that are likely to be of interest to the user without requiring the user to perform an active search; (2) to enhance targeting

of advertising to a particular user's interests; and (3) to augment active search queries based on user profile information; and the like.

8. Similarity-measure augmentation of IS-based systems.

One implementation provides a similarity augmentation for an IS-based system. People have the ability to recognize the similarity or dissimilarity of subfields of a field, whether the field is subject matter or some other distinction (geographic information, activity, and the like). This recognition is not based on any relationships between terms in the fields of subfields (although such similarity may be found in some cases). This ability allows people to recognize metaphor, in which text about one field includes language from a very different field (e.g., sport and war); change of subject (and degree of change of subject), when an answer is not responsive to a question; and so forth.

The similarity augmentation for IS-based systems provides an additional IS space in which the similarity of each pair of topics (subject matter fields or other distinction) is denoted by a rating on a numerical scale of fixed range, as terms and fields are rated in IS spaces. For example, using 4 subject matter fields and a 0 to 1 scale for similarity values, one might have:

	Physics	Finance	Business	Poetry
Physics	1.0	0.2	0.0	0.0
Finance	0.2	1.0	0.8	0.0
Business	0.0	0.8	1.0	0.0
Poetry	0.0	0.0	0.0	1.0

Given two texts, the IS system calculates the IS of each text. The similarity values are used to numerically weight the difference between the fields not in common in the two spectra. For example, if Document A has a spectrum of 5.6 on Business and Document B has a spectrum of 4.3 on Physics, the Physics-Finance similarity of 0.2 is a weighting factor in calculating the distance between Document A and Document B: The distance between Document A and Document B is multiplied by (1.0 - 0.2) = 0.8. Other formulas for this weighting can be used.

This similarity augmentation is combined with an IS-based system to add an additional layer of recognition of attributes of the text. The augmentation computes degree of disparateness of two documents.

9. IS-based query assessor.

One implementation provides an IS-based query assessor.

Conventional information retrieval and search engines evaluate a query by producing a set of "search terms." Such search terms are words and phrases to be searched for in documents. The IS-based query assessor produces a multiple-aspect IS, which may be linked, which measures the content of the query, in the manner described above. This numerical representation of the query's "aboutness" can be provided to an IS-based search engine.

10. Depth-of-content assessor.

One implementation provides a depth-of-content assessor.

It is frequently desirable to be able to automatically recognize whether the content of a text is at a superficial level of a field, employing only the broad concepts of the field, or whether the content deals with more subtle, fine-grained distinctions of a field, such as might be found in college-level, graduate-level, or professional-level literature. Persons familiar with the subject matter of a field can recognize this characteristic, just as they can recognize to what field a text is relevant. The depth-of-content assessor provides this capability.

The fields and subfields of a type of content form a hierarchy. Distinctions within successively lower fields in the hierarchy represent successively finer and finer distinctions within the subject. Therefore, a text's set of spectra for a type of content represents the amount of content at greater and greater *conceptual* depth, not merely numerical depth, in the hierarchy.

A term that is part of a field, but not part of any subfield of that field, does not contribute to the spectrum of the text within the field. Consider text containing a term t, where t appears within a field F, but not within any of the subfields of F. In this example, the field F contributes to the IS for the text, but the subfields of F do not contribute to the IS for the text.

As a result, a text with terms only at level *n* of a hierarchy has spectra only down to level *n* of it. In this way, the depth-of-content assessor automatically measures the depth of conceptual content of the text.

For example, "atom" is a term referring to a concept in the field of physics, but not in the field of subatomic physics. A text with only such terms as this would not have an IS within the field of particle physics, although it would within physics. By contrast, a document with the terms "Schrodinger's equation", "S-matrix", and "Feynman diagram" would have spectra within the subfields of physics representing this deep content.

11. Text-quality based retrieval system.

One implementation provides a text quality based retrieval system.

The quality of a text can be determined by measuring different aspects of the text, such as accuracy, authoritativeness, lack of bias, and depth of content. For example, documents satisfying a query can be retrieved based on the number of documents that cite them.

The depth-of-content assessor described above provides a measure of depth of content of a text. The text quality based retrieval system retrieves documents based on this measure of depth of content of each document. For example, the text quality based retrieval system can be used with an IS-based search engine can select documents to be retrieved by matching the depth of content of the documents with the depth of content of a search request.

12. IS-based post-retrieval filter of any list of documents.

One implementation provides an IS-based post-retrieval filter.

Conventional retrieval systems retrieve many documents, only a portion of which are relevant to the user's request. It often happens that many relevant documents are retrieved, but a very large number of irrelevant ones are retrieved as well. This renders the retrieved set of documents very difficult to use.

The IS-based post-retrieval filter takes as input all documents retrieved by any retrieval system, calculates the IS of each, and arranges the documents in closest-profile-first order. When used in conjunction with a primary information retrieval system of any design, the IS-based post-retrieval filter provides a complementary relevance ranking enhancement.

The IS-based post-retrieval filter analyzes the query and calculates an IS for the query. The primary information retrieval system uses the query to retrieve an initial set of documents. The IS-based post-retrieval filter calculates an IS for each document in the initial set. The IS for the documents is compared to the IS of the query, and the documents in the initial set are reordered based on the IS distance between the documents and the query. This IS-based relevance ranking can be weighted to allow for the inclusion of other relevance measures and to suit the needs of administrators of the primary information retrieval system.

13. IS-based query expander.

One implementation provides an IS-based query expander.

Existing retrieval system often add to a user's query other terms, words or phrases. This process is known as *query expansion.*

In an IS hierarchy, each space in the hierarchy has a set of terms that are rated as highly relevant to one or more of the subfields of that space. The IS-based query expander calculates the IS of the user's query, and produces as output all terms from the content hierarchies of the system that are highly relevant to any field of hierarchy in which the query has content. These output terms are added to the user's query. The expanded query is then provided to a search engine of any design for document searching.

14. Search engine based on IS-query expansion and IS-post-filtering.

The IS-based query expander and the IS-based post-retrieval filter can be used to "sandwich" a search engine of any type to improve its capabilities. The query expander expands queries prior to searching as described above. The filter ranks the results as described above.

15 IS-based intelligent crawler.

One implementation provides an IS-based intelligent crawler.

Existing network retrieval devices, commonly known as "crawlers" or "spiders", build a list of links to documents or "pages" on the network, retrieve those pages, add the links from the retrieved pages to the list of links, and so on. Various conventional techniques have been used to enhance this algorithm, including analyzing the statistics of links between pages and analyzing which pages are most popular with users.

The IS-based intelligent crawler uses the information spectrum of a page to change the order of the links to be followed, mimicking a person choosing books in a

library by following bibliographic references in each book. The crawler begins with a request for information (which may include one or more documents or parts of documents). As a page is retrieved by the crawler, its IS is calculated. When the distance between a page and the request exceeds a predetermined threshold distance, set by the user or an administrator, its links are not added to the list of links to be followed or added at a lower priority.

Thus the IS-based intelligent crawler device emulates the logic of a person researching a topic by following bibliographic citations in a library. As the person examines a cited work, he may recognize references as similar in content to the topic, in which case he follows the citations in the reference. Alternatively, he may notice that the content is getting far from the topic, and stop following a chain of references. IS-based crawling embodies this ability to recognize when content is becoming more dissimilar to the original information need.

16. IS-based information search on a network.

One implementation provides an intelligent crawler search.

Any information system in which documents are linked to each other via explicit references forms a "web" of documents in which one document is linked to another by an author's recommendation. The Internet is the largest example of such a structure. When an author of a document puts a hypertext link in it, he is giving the reader a "See also" recommendation.

The intelligent crawler search provides a way to navigate such document networks or "webs," including the Internet. At any time, a user can select any combination of portions of documents or entire documents and indicate "Find information sources like this." The intelligent crawler search returns to the user other sources on the network that are similar in content to the indicated text.

A user provides a query, which can be a word, phrase, or one or more documents or portions of documents. The query is then used as the starting point for the IS-based intelligent crawler described above. The documents found by the intelligent crawler are provided to the user.

17. IS-based intrusion detector.

One implementation provides an IS-based intrusion detector.

IS hierarchies are not limited to text. Any set of distinctions a person can make, for which there are automatically recognizable exemplars, can be used for an IS hierarchy. For example, an IS space can include terms that are ways of referring to symptoms and fields that are the diagnoses. Such an IS space can form the basis of a system that can function as an expert, such as a help system, a troubleshooting system, a diagnostic system, and the like. A problem report can serve as the query, and troubleshooting guides can serve as the documents to be found.

The IS-based intrusion detector is an IS-based system in which the "fields" are types of computer security intrusions, and the "terms" are detectable sequences of commands received from a source at the site of a computer attached to a network. Requests and/or commands are monitored, and are used to calculate an ongoing threat profile. In most cases a single command or request is not enough to define a threat profile above the threshold at which action should be taken, just as in most cases a single term in a document does not produce a content spectrum with significant values. However, as the stream of commands and requests of a user continues it creates an aggregate threat profile. When a component of a threat profile achieves an administrator-defined threshold, a threat is detected, and alerting or other measures are taken. The system may also match the existing threat profile against known prior intrusions, detecting a threat when the threat profile becomes close enough to a known previous intrusion or intrusion attempt.

18. IS-based search engine for other resources

The items examined during a search are not limited to documents or web pages. For example, the IS-based search engines described above can be used to monitor the content within chat rooms, and to alert the user when a subject is under discussion. Other examples include the contents of books for sale, and the content of advertisements.

19. An IS-based spam filter.

"Spam" is the customary name given to unwanted email and notifications. It is also applied to Web pages with misleading terms, that do not reflect the actual content of the page, so that automatic document-finding programs or Web-page finding programs, such as network "crawlers" (often referred to as "spiders") will find these pages although the person to whom they are returned is not attempting to find pages with this kind of content.

The IS-based spam filter recognizes this kind of email or Web page by calculating the IS of the document in question, as described above. Terms that are unrelated to the content of the page result in an IS with high values on at least one field other than the one or ones that reflect the actual document's content. If a user has specified his or her information interests in a user profile as described above, a "spam" document will have a greater distance from the user's profile than documents with similar content. The IS-based spam filter then redirects this document as the user or an administrator determines: rejection, re-direction to a special location or account, and the like.

20. Multiple Languages

The implementations described above are not limited to searches in a single language. A hierarchy can include terms and fields in multiple languages. An IS-based system using a hierarchy with terms from more than one language can find items with similar aggregate content spectra even when those items are in different languages. Further, a given item can include text in multiple languages. The above-described implementations function equally well with such items.

The invention can be implemented in digital electronic circuitry, or in computer hardware, firmware, software, or in combinations of them. Apparatus of the invention can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor; and method steps of the invention can be performed by a programmable processor executing a program of instructions to perform functions of the invention by operating on input data and generating output. The invention can be implemented advantageously in one or more computer programs that are executable on a programmable system including at least one programmable processor coupled to receive data and instructions from, and to transmit data and instructions to, a data storage system, at least one input device, and at least one output device. Each computer program can be implemented in a high-level procedural or object-oriented programming language, or in assembly or machine language if desired; and in any case, the language can be a compiled or interpreted language. Suitable processors include, by way of example, both general and special purpose microprocessors. Generally, a processor will receive instructions and data from a read-only memory and/or a random access memory. Generally, a computer will include one or more mass storage devices for storing data files; such devices include magnetic disks, such as internal hard disks and removable disks; magneto-optical disks; and optical disks. Storage devices suitable for tangibly embodying computer program instructions and data include all

forms of non-volatile memory, including by way of example semiconductor memory devices, such as EPROM, EEPROM, and flash memory devices; magnetic disks such as internal hard disks and removable disks; magneto-optical disks; and CD-ROM disks. Any of the foregoing can be supplemented by, or incorporated in, ASICs (application-specific integrated circuits).

A specific embodiment of a method and apparatus for providing a search engine according to the present invention has been described for the purpose of illustrating the manner in which the invention is made and used. It should be understood that the implementation of other variations and modifications of the invention and its various aspects will be apparent to-one skilled in the art, and that the invention is not limited by the specific embodiments described. Therefore, it is contemplated to cover the present invention and any and all modifications, variations, or equivalents that fall within the true spirit and scope of the basic underlying principles disclosed and claimed herein.

WHAT IS CLAIMED IS:

1. A method comprising:

segmenting a judgment matrix into a plurality of information sub-matrices where each submatrix has a plurality of classifications and a plurality of terms relevant to each classification; evaluating a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information submatrices;

calculating an information spectrum for a first document based upon at least some of the plurality of terms;

calculating an information spectrum for a second document based upon at least some of the plurality of terms; and

identifying the second document as relevant to the first document based upon a comparison of the calculated information spectrums.

2. The method of claim 1, wherein the step of segmenting a judgment matrix further comprises dividing each information submatrix of the information sub-matrices into a set of columns, where each column is a classification of the information sub-matrix.

3. The method of claim 1 further comprising selecting the plurality of terms based up a relevance of each term of the plurality of terms to at least some of the classifications of the information sub-matrices.

4. The method of claim 1, wherein evaluating a relevance further comprises assigning a numerical indicia of relevance having a range of between zero and at least two.

5. The method of claim 4, wherein calculating an information spectrum for the first and second documents further comprises determining a log average among the numerical indicia of relevance of the terms of each classification.

6. The method of claim 1, wherein identifying further comprises determining a distance between the information spectrum of the first document and the information spectrum of the second document.

7. The method of claim 1 further comprising indicating that the first and second documents are relevant to each other.

8. The method of claim 7 further comprising using the calculated information spectrum of at least one of the first and second documents as a search request.

9. The method of claim 1 further comprising zooming in on a portion of a document information spectrum.

10. The method of claim 9 further comprising determining that the first and second documents have a wide spectra with significant content in a field F of a term.

11. The method of claim 10 further comprising measuring the first and second document using a sub-engine for field F.

12. The method of claim 1, wherein the first document is a user interest profile describing one or more interests of a user and the second document is part of a text stream, further comprising:

directing the first document to the user when the second document is identified as relevant to the first document.

13. The method of claim 1, wherein the first document is a profile of a user's job-related information needs, further comprising:

denying the user access to the second document when the second document is not identified as relevant to the first document.

14. The method of claim 1, wherein the first document is a profile of objectionable content and the second document includes one or more document hosted by a web site, further comprising:

denying access to the web site when the second document is identified as relevant to the first document.

15. The method of claim 1, wherein:

calculating comprises calculating a depth-of-content measure for a document based on the lowest sub-matrix in which a term in the document appears; and

identifying comprises comparing the depth-of-content measures of the first and second documents.

16. The method of claim 1, further comprising:

calculating an information spectrum for a third document based upon at least some of the plurality of terms; and

ranking the second and third documents in order of degree of relevance to the first document.

17. The method of claim 1, wherein the first document is a search request and the second document includes a hyperlink to a third document, further comprising:

calculating an information spectrum for a third document when the second document is identified as relevant to the first document.

18. The method of claim 1, wherein the terms are computer instructions, the first document is a threat profile, and the second document is a set of instructions received from a source, further comprising:

detecting a threat when the second document is identified as relevant to the first document.

19. The method of claim 1, wherein the first document is a user interest profile describing one or more interests of a user and the second document is directed to the user, further comprising:

redirecting the second document when the second document is identified as not relevant to the first document.

20. A method for augmenting a document, comprising:

segmenting a judgment matrix into a plurality of information sub-matrices where each sub-matrix has a plurality of classifications and a plurality of terms relevant to each classification;

evaluating a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information sub-matrices; and

adding at least one of the terms to the document based on the relevance of the terms.

21. A method for evaluating a text having a plurality of terms, comprising:

calculating a first information spectrum for the text in a first judgment matrix having a plurality of first vector spaces arranged in a hierarchy;

calculating a second information spectrum for the text in a second judgment matrix; and

combining the first and second information spectra to produce an information spectrum for the text.

22. The method of claim 21, further comprising:

calculating a document distance between the information spectrum for the text and a further information spectrum for a further text; and

identifying the further text as relevant to the text when the document distance falls below a predefined threshold.

23. The method of claim 22, wherein calculating a second information spectrum comprises:

identifying as contributing terms any terms that appear in at least one of the first hierarchical vector spaces;

identifying as associated terms any terms that are associated with one or more of the contributing terms; and

calculating a linked information spectrum for the text in a second judgment matrix using each associated term.

24. The method of claim 23, further comprising:

calculating a document distance between the information spectra for the text and further information spectra for a further text; and

identifying the further text as relevant to the text when the document distance falls below a predefined threshold.

25. A computer program product, tangibly stored on a computer-readable medium, comprising instructions operable to cause a programmable processor to:

segment a judgment matrix into a plurality of information sub-matrices where each submatrix has a plurality of classifications and a plurality of terms relevant to each classification; evaluating a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information submatrices;

calculate an information spectrum for a first document based upon at least some of the plurality of terms;

calculate an information spectrum for a second document based upon at least some of the plurality of terms; and

identify the second document as relevant to the first document based upon a comparison of the calculated information spectrums.

26. The computer program product of claim 25, wherein the instructions operable to cause a programmable processor to segment a judgment matrix further comprises instructions operable to cause a programmable processor to divide each information submatrix of the information sub-matrices into a set of columns, where each column is a classification of the information sub-matrix.

27. The computer program product of claim 25 further comprising instructions operable to cause a programmable processor to select the plurality of terms based up a relevance of each term of the plurality of terms to at least some of the classifications of the information sub-matrices.

28. The computer program product of claim 25 wherein the instructions operable to cause a programmable processor to evaluate a relevance further comprises instructions operable to cause a programmable processor to assign a numerical indicia of relevance having a range of between zero and at least two.

29. The computer program product of claim 28, wherein the instructions operable to cause a programmable processor to calculate an information spectrum for the first and second documents further comprises instructions operable to cause a programmable processor to determine a log average among the numerical indicia of relevance of the terms of each classification.

30. The computer program product of claim 25 wherein the instructions operable to cause a programmable processor to identify further comprises instructions operable to cause a programmable processor to determine a distance between the information spectrum of the first document and the information spectrum of the second document.

31. The computer program product of claim 25 further comprising instructions operable to cause a programmable processor to indicate that the first and second documents are relevant to each other.

32. The computer program product of claim 31 further comprising instructions operable to cause a programmable processor to use the calculated information spectrum of at least one of the first and second documents as a search request.

33. The computer program product of claim 25 further comprising instructions operable to cause a programmable processor to zoom in on a portion of a document information spectrum.

34. The computer program product of claim 33 further comprising instructions operable to cause a programmable processor to determine that the first and second documents have a wide spectra with significant content in a field F of a term.

35. The computer program product of claim 34 further comprising instructions operable to cause a programmable processor to measure the first and second document using a sub-engine for field F.

36. The computer program product of claim 25, wherein the first document is a user interest profile describing one or more interests of a user and the second document is part of a text stream, further comprising instructions operable to cause a programmable processor to:

direct the first document to the user when the second document is identified as relevant to the first document.

37. The computer program product of claim 25, wherein the first document is a profile of a user's job-related information needs, further comprising instructions operable to cause a programmable processor to:

deny the user access to the second document when the second document is not identified as relevant to the first document.

38. The computer program product of claim 25, wherein the first document is a profile of objectionable content and the second document includes one or more document hosted by a web site, further comprising instructions operable to cause a programmable processor to:

deny access to the web site when the second document is identified as relevant to the first document.

39. The computer program product of claim 25, wherein:

the instructions operable to cause a programmable processor to calculate comprises instructions operable to cause a programmable processor to calculate a depth-of-content measure for a document based on the lowest sub-matrix in which a term in the document appears; and

the instructions operable to cause a programmable processor to identify comprises instructions operable to cause a programmable processor to compare the depth-of-content measures of the first and second documents.

40. The computer program product of claim 25, further comprising instructions operable to cause a programmable processor to:

calculate an information spectrum for a third document based upon at least some of the plurality of terms; and

rank the second and third documents in order of degree of relevance to the first document.

41. The computer program product of claim 25, wherein the first document is a search request and the second document includes a hyperlink to a third document, further comprising instructions operable to cause a programmable processor to:

calculate an information spectrum for a third document when the second document is identified as relevant to the first document.

42. The computer program product of claim 25, wherein the terms are computer instructions, the first document is a threat profile, and the second document is a set of instructions received from a source, further comprising instructions operable to cause a programmable processor to:

detect a threat when the second document is identified as relevant to the first document.

43. The computer program product of claim 25, wherein the first document is a user interest profile describing one or more interests of a user and the second document is directed to the user, further comprising instructions operable to cause a programmable processor to:

redirect the second document when the second document is identified as not relevant to the first document.

44. A computer program product for augmenting a document, comprising instructions operable to cause a programmable processor to:

segment a judgment matrix into a plurality of information sub-matrices where each sub-matrix has a plurality of classifications and a plurality of terms relevant to each classification;

evaluate a relevance of each term of the plurality of terms with respect to each classification of each information sub-matrix of the information sub-matrices; and

add at least one of the terms to the document based on the relevance of the terms.

45. A computer program product for evaluating a text having a plurality of terms, comprising instructions operable to cause a programmable processor to:

calculate a first information spectrum for the text in a first judgment matrix having a plurality of first vector spaces arranged in a hierarchy;

calculate a second information spectrum for the text in a second judgment matrix; and

combine the first and second information spectra to produce an information spectrum for the text.

46. The computer program product of claim 45, further comprising instructions operable to cause a programmable processor to:

calculate a document distance between the information spectrum for the text and a further information spectrum for a further text; and

identify the further text as relevant to the text when the document distance falls below a predefined threshold.

47. The computer program product of claim 45, wherein the instructions operable to cause a programmable processor to calculate a second information spectrum comprises instructions operable to cause a programmable processor to:

identify as contributing terms any terms that appear in at least one of the first hierarchical vector spaces;

identify as associated terms any terms that are associated with one or more of the contributing terms; and

calculate a linked information spectrum for the text in a second judgment matrix using each associated term.

48. The computer program product of claim 47, further comprising instructions operable to cause a programmable processor to:

calculate a document distance between the information spectra for the text and further information spectra for a further text; and

identify the further text as relevant to the text when the document distance falls below a predefined threshold.

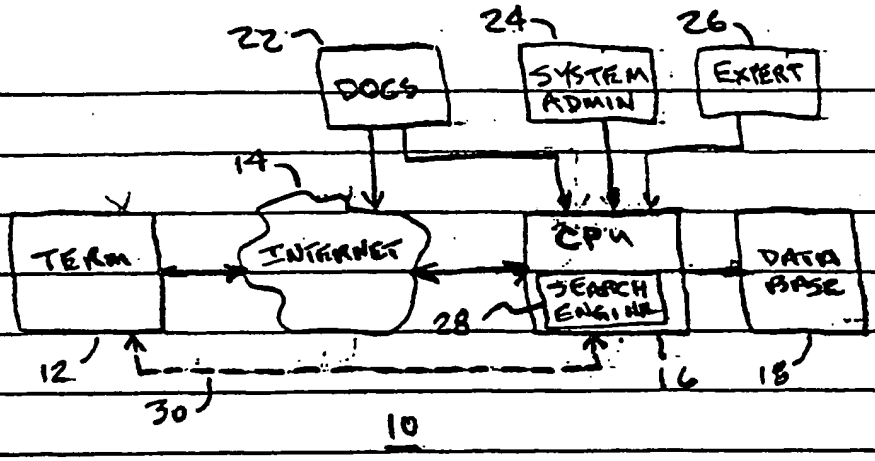

22
24
26
DOGS
SYSTEM ADMIN
EXPERT
14
TERM
INTERNET
CPU
SEARCH ENGINE
DATA BASE
28
12
30
16
18
10

FIG. 1

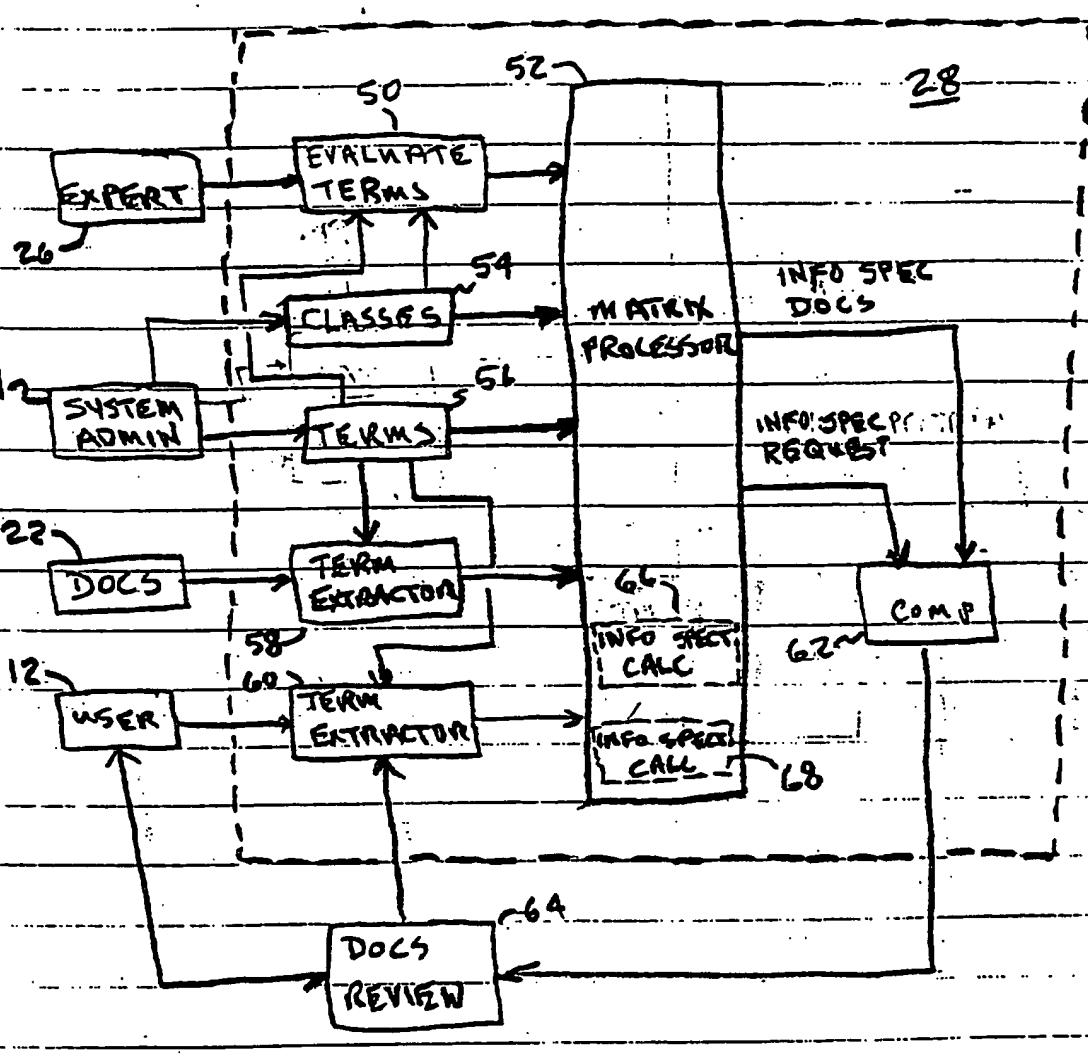

28
50
52
EVALUATE TERMS
EXPERT
26
54
CLASSES
MATRIX PROCESSOR
INFO SPEC DOCS
24
SYSTEM ADMIN
56
TERMS
INFO SPEC REQUEST
22
DOCS
TERM EXTRACTOR
66
COMP
INFO SPEC CALC
62
58
12
60
USER
TERM EXTRACTOR
INFO SPEC CALL
68
64
DOCS REVIEW

FIG. 2

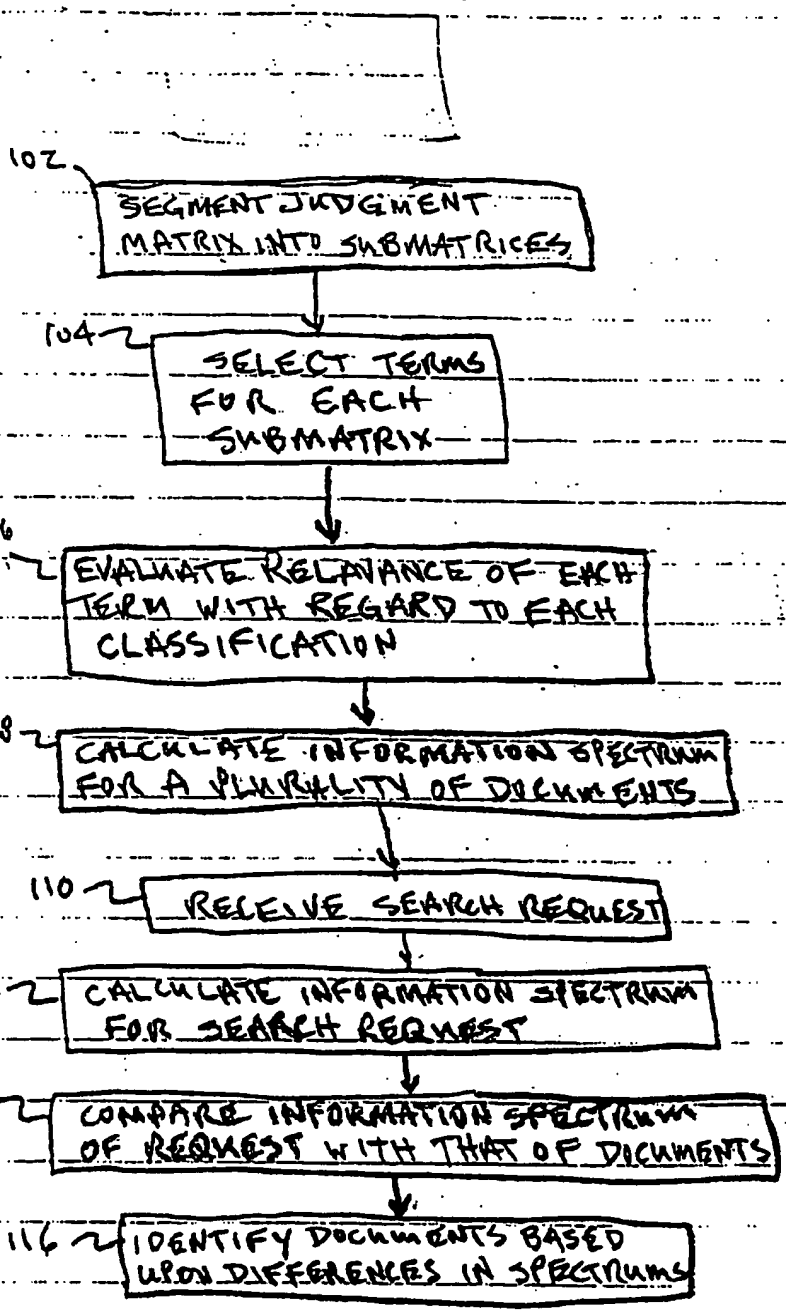

102
SEGMENT JUDGMENT MATRIX INTO SUBMATRICES
104
SELECT TERMS FOR EACH SUBMATRIX
106
EVALUATE RELAVANCE OF EACH TERM WITH REGARD TO EACH CLASSIFICATION
108
CALCULATE INFORMATION SPECTRUM FOR A PLURALITY OF DOCUMENTS
110
RECEIVE SEARCH REQUEST
112
CALCULATE INFORMATION SPECTRUM FOR SEARCH REQUEST
114
COMPARE INFORMATION SPECTRUM OF REQUEST WITH THAT OF DOCUMENTS
116
IDENTIFY DOCUMENTS BASED UPON DIFFERENCES IN SPECTRUMS

FIG. 3

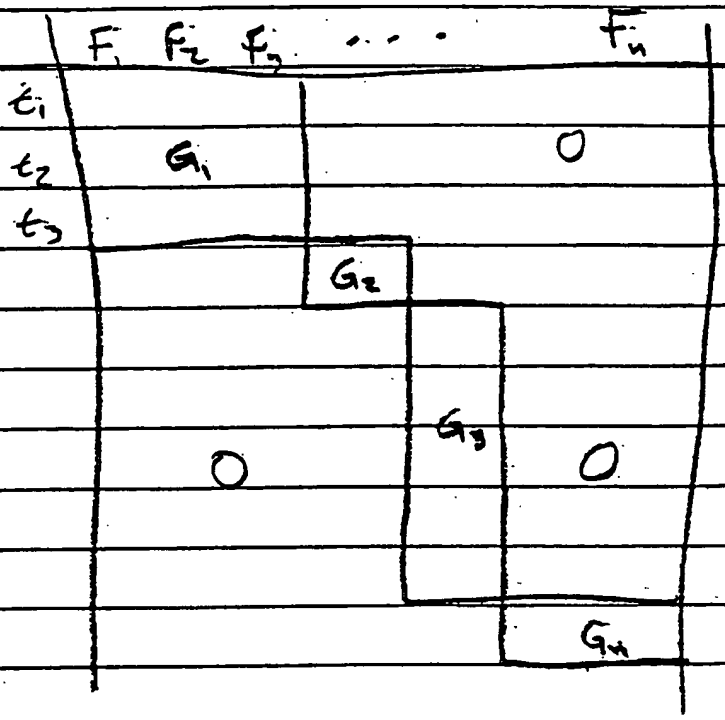


FIG. 4

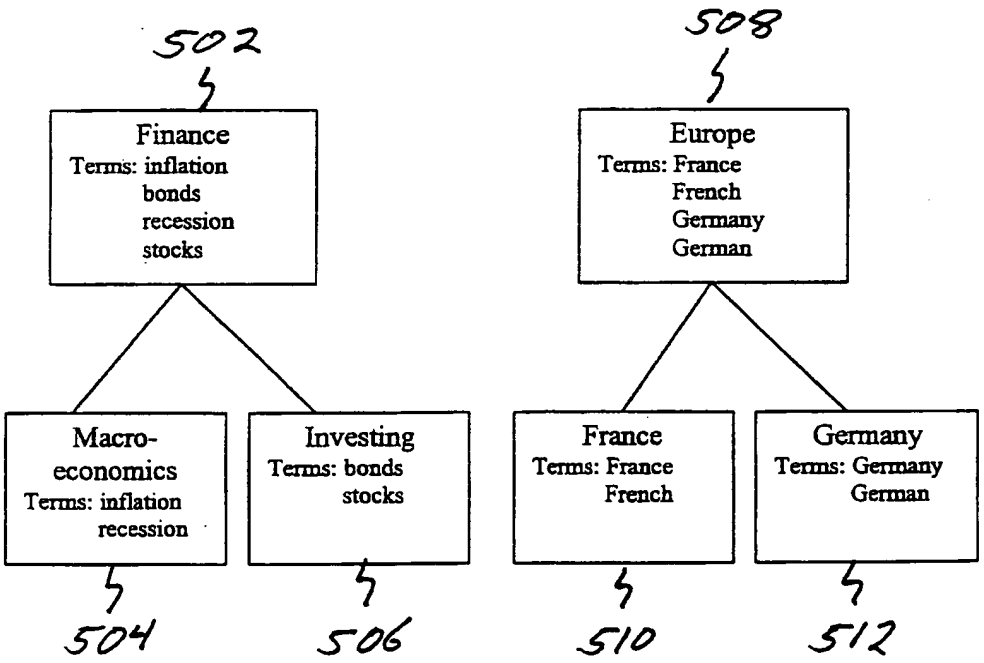

502
Finance
Terms: inflation
bonds
recession
stocks
508
Europe
Terms: France
French
Germany
German
Macro-
economics
Terms: inflation
recession
Investing
Terms: bonds
stocks
France
Terms: France
French
Germany
Terms: Germany
German
504
506
510
512

FIG. 5

INTERNATIONAL SEARCH REPORT

Ir .ational Application No
PCT/US 00/12344

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
WPI Data, EPO-Internal, INSPEC, IBM-TDB, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TAO GU: "An Automated Method of Document Classification and Retrieval by Means of Term Subspaces" CONFERENCE PROCEEDINGS IEEE SOUTHEASTCON '82, 4 - 7 April 1982, pages 26-33, XP000955681 Sandestin, FL, USA page 26, left-hand column, line 1 -page 28, left-hand column, line 40	1,20,21, 25,44,45
A	US 5 857 179 A (ADLER MARK R ET AL) 5 January 1999 (1999-01-05) column 2, line 20 -column 3, line 50 abstract	1,20,21, 25,44,45

-/--

[X] Further documents are listed in the continuation of box C. [X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 September 2000	04/10/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Deane, E

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/12344

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 98 58344 A (DIALOG CORP) 23 December 1998 (1998-12-23) page 11, line 12 -page 17, line 20; figures 1,2,3A,3B	1,20,21, 25,44,45

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

h ational Application No
PCT/US 00/12344

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5857179 A	05-01-1999	NONE	
WO 9858344 A	23-12-1998	AU 8373798 A	04-01-1999
		EP 0996927 A	03-05-2000